APPENDIX C

General Data Quality Issues

GENERAL DATA QUALITY ISSUES

To add perspective to the types of data quality issues encountered while working with the occurrence data, some of the most common concerns are outlined below. While this is far from a complete list of all the quality issues encountered, it should provide some idea of the various complexities and complications involved in preparing the data.

Every State database contained unique data elements or a unique treatment of common elements. Even after initial screening and conversion, factors were always uncovered during data analysis. These were resolved in consultation with the State or data source. As a general rule, when errors or ambiguities in various data elements could not be resolved, the observation was eliminated from the analysis to avoid aberrant results. These observations made up only a very small portion of the large number of records included in most data sets.

Structure of the Data

Most of the data sets are "vertically" designed, meaning there is one column for the system identification number, one column for the date sampled, one column for the name of the contaminant analyzed, one column for the results, etc. Other data sets are "horizontally" designed, meaning there is one column for the system identification number and multiple columns for the contaminants analyzed (each contaminant has a separate column). The results for each contaminant are thus displayed along a single data row under the appropriate contaminant heading. A horizontal data structure is far more difficult to analyze in the manner necessary for this study and requires extensive processing to transform into a more suitable vertical structure.

File Format

Files which were submitted in CSV (comma separated values) format were problematic if proper care was not taken to exclude or modify data fields (usually text fields) which might contain commas within them (for example, a field holding the address of a water system might contain one or more commas). This field would often separate into two or more columns as the computer read the file. Unless every observation contained the same number of text commas within each field, the following data fields would no longer line up by column between observations. Making the necessary adjustments to allow the computer to properly recognize and process the correct data fields is a tedious and time intensive task, even with the use of specialized programs.

Multiple Data Sets

The size and number of data sets required to make up a complete occurrence database for a State can vary considerably. While some States maintain one database for all water system monitoring results, it is not unusual for a State to keep a number of databases for various data subsets (e.g., a separate database for each of the contaminant groups – IOCs, SOCs, and VOCs). Other States have very elaborate subsets consisting of multiple (10 or more) separate data sets for all of the compliance data for the State. Sometimes each subset will have a unique format and structure and require significant formatting before it is compatible with the rest of the State data. Many

States have an individual data manager for each of the contaminant groups and coordinating data transfer requires communication with more than one contact person.

Contaminant Codes

Analytes were identified by a variety of codes including EPA codes, CAS numbers, STORET codes (which might have multiple codes for a given contaminant), State-specific or laboratory-specific codes, or by chemical name. In every case where a contaminant was identified by any system other than EPA codes, the proper EPA contaminant designation was found and added to the record.

Some States have special coding systems for contaminants that are covered under the same analytical method. One of the most common systems summarizes the results of a single method with an 'ND' or zero for all contaminants not detected and individual observations only for those contaminants with a positive result. This system is best illustrated using the 21 VOCs covered under method 502.2 as an example. If none of the 21 VOCs were detected, a State might enter '21 VOCs' in the contaminant column and '0' in the results column of a single observation. If one or more of the 21 VOCs were detected, these individual contaminants would be entered in the contaminant column and the value they were detected at would be entered in the results column of individual observations. It is assumed that every contaminant not recorded individually was tested for and not detected. Although the example above may seem straightforward, the rules can become complicated. It is not unusual to find caveats in the database which need to be further defined by the State before analysis can begin. This is done differently between States, and it is often not entirely consistent within a State data set.

Reporting non-Detections

There are numerous ways that States report non-detections in their data systems. Some States report a value of zero when a contaminant is not detected, others have a column in which they enter a less-than sign and then enter the method detection limit or reporting level in the results column. A third common method for reporting no detects is the inclusion of a separate text variable that will read 'ND'. A result value of ND would be converted to zero for the purposes of this study, while less-than values were kept as they are to provide a sense of the various MRL/MDLs.

Units

States differ in the units they use to enter contaminant concentration results and, at times, the reporting units change within a State. The most common unit for reporting results is in milligrams per liter, but at times only the SOCs are reported in milligrams per liter and the rest of the database will be in micrograms per liter. In some cases, units are not identified in a data set.

Other Data

Most of the databases coming from the States are not designed for data analysis and are generally "electronic filing cabinets." The databases contain special sampling data, raw water data, compliance data, and other sampling data. To analyze only the compliance data, it is necessary to understand the coding system the State uses to distinguish the compliance sampling from the other types of sampling. Most coding systems are unique to each State.